Cross-lingual Transfer of Correlations between Parts of Speech and Gaze Features

Cross-lingual Transfer of Correlations between Parts of Speech and Gaze Features

Maria Barrett
Centre for Language Technology
University of Copenhagen
Njalsgade 136
2300 Copenhagen S, Denmark
barrett@hum.ku.dk

Frank Keller
School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB, UK
keller@inf.ed.ac.uk

Anders Søgaard
Dpt. of Computer Science
University of Copenhagen
Sigurdsgade 41
2200 Copenhagen N, Denmark
soegaard@di.ku.dk

Abstract

Several recent studies have shown that eye movements during reading provide information about grammatical and syntactic processing, which can assist the induction of NLP models. All these studies have been limited to English, however. This study shows that gaze and part of speech (PoS) correlations largely transfer across English and French. This means that we can replicate previous studies on gaze-based PoS tagging for French, but also that we can use English gaze data to assist the induction of French NLP models.

1 Introduction

The eye movements during normal, skilled reading are known to reflect the processing load associated with reading. Recently, eye movement data has been integrated into natural language processing models for weakly supervised part-of-speech (PoS) induction (Barrett et al., 2016), sentence compression (Klerke et al., 2016), supervised PoS tagging (Barrett and Søgaard, 2015a), and supervised parsing (Barrett and Søgaard, 2015b).

Barrett et al. (2016) used eye movements from the English portion of a large eye tracking corpus, the Dundee corpus (Kennedy et al., 2003), for weakly supervised PoS induction for English, obtaining significant improvements over a baseline without gaze features. They used a second-order hidden Markov Model, which was type-constrained by Wiktionary dictionaries for their experiments. These results suggest an approach to weakly supervised PoS induction using only a dictionary and eye movement data. Such an approach would be applicable for low-resource languages, for which it is difficult to find professional annotators.

The present study further explores to which extent native readers' processing of PoS generalizes across related languages. We use a similar model as Barrett et al. (2016), but perform cross-lingual experiments with both the French and the English portion of the Dundee Corpus.

Contribution This is to the best of our knowledge the first study to explore how the eye movements of native readers that inform PoS models generalize from one language to another. We also introduce a new resource for studying the relation between grammatical class and eye movements in French: we provide PoS annotation for most of the French Dundee Corpus by aligning it with the morphosyntactic annotation of the French Treebank (Abeillé et al., 2003).

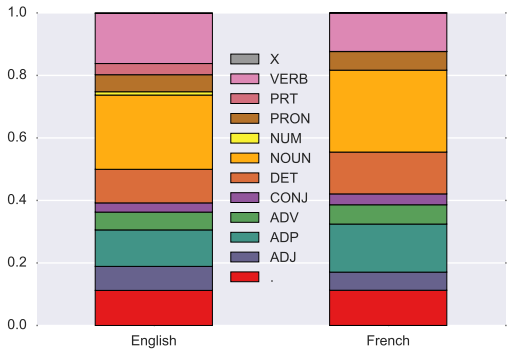
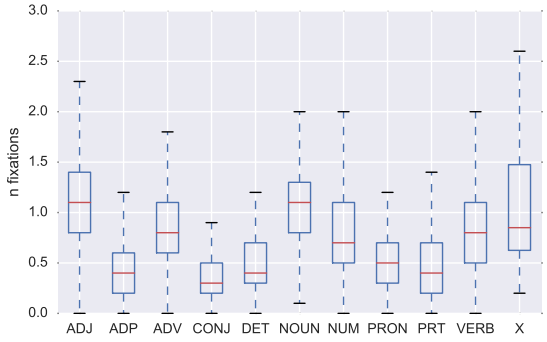
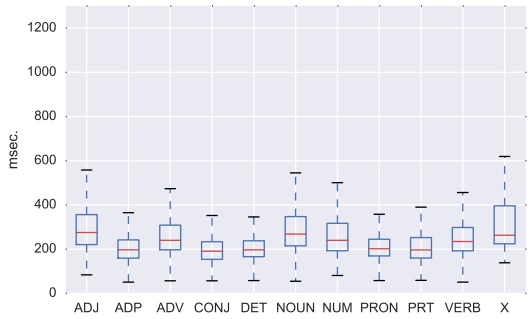


Figure 1: Distribution of PoS in the English and French training sets.

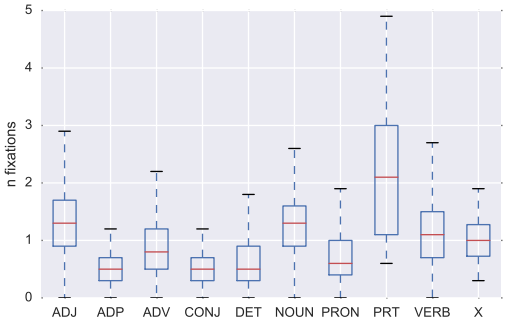
Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1330–1339, Osaka, Japan, December 11-17 2016.



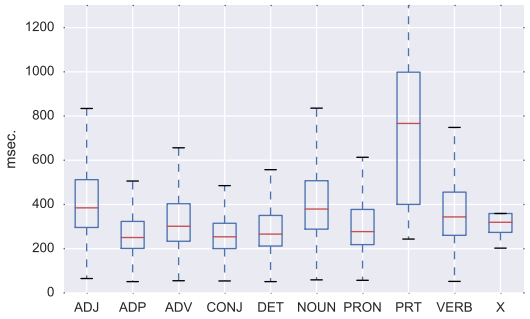
(a) English number of fixations



(b) English First pass duration



(c) French number of fixations



(d) French first pass duration

Figure 2: Two reading measures across PoS class computed on the English and French training sets.

2 Data preparation

The data used for this experiment is the English and French portions of the Dundee Corpus (Kennedy et al., 2003). The Dundee Corpus is the largest available eye movement corpus by token count. For English and French, 10 native speakers of each language read 20 newspaper articles from either *The Independent* (English) or *Le Monde* (French). The corpus comprises around 50,000 tokens per language.

For both the English and the French part of the Dundee Corpus, the original tokenization follows the visual units of the text, and contractions and punctuation are attached to the word whose visual unit they belong to. For instance, *s'entendre* or *rappelle-t-il* are one token in the French Dundee Corpus but two and five, respectively, in the French Treebank. In the English Dundee Corpus, *don't!* is one token, but three in the Dundee Treebank. As a result, eye movement measures are only available for the entire visual unit. We address this issue by duplicating the eye movement measures for all treebank tokens that comprise a Dundee token (i.e., a visual unit). This is the same approach Barrett et al. (2016) used. As a result, the number of tokens increases in the PoS-tagged version of the Dundee Corpus; also, some tokens are associated with eye movement measures that reflect the processing of several tokens.

For English, the treebank tokenization leads to 13.8% increase of tokens to 58,599 tokens. For French, the treebank tokenization leads to an 17.7% increase on token count to 56,683 tokens. For the English training set, 76% of all Dundee Corpus tokens are mapped to one treebank token. The same goes for 62% of the Dundee Corpus tokens for French.

2.1 English

The Dundee Treebank (Barrett et al., 2015) is a recent manual, syntactic annotation layer for the English portion of the Dundee Corpus following the Universal Dependency formalism. For evaluation, we use the PoS labels from this resource. We mapped the Penn Treebank tagset used in the Dundee Treebank

automatically to the Universal PoS tag set (Petrov et al., 2011).

The split into training, development, and test set for the English Dundee corpus is identical to the splits used by Barrett et al. (2016), with 80% of the tokens for training and 10% of the tokens for development and testing, respectively, without splitting up sentences. This split results in 46,879 tokens in 1,896 sentences for training, 5,868 tokens in 230 sentences for development, and a test set of 5,832 tokens in 241 sentences.

2.2 French

The text for the French part of the Dundee Corpus is originally from the French Treebank version 1.4 (Abeillé et al., 2003) and we re-aligned the two corpora for this experiment. We first manually identified the relevant subset of the French Treebank (which is discontinuous). A small part (2,518 tokens equivalent of 5.31% of the French Dundee tokens) of the Dundee Corpus could not be found by manual search in the French Treebank and was therefore omitted from the experiment. Only entire sentences were removed. The morphosyntactic annotation of the French Treebank was semi-manually aligned with the Dundee Corpus by a set of heuristic rules and by manually fixing all exceptions. Due to tokenization inconsistencies in both the French Treebank and the Dundee Corpus, manual intervention was required.

For French there are some treebank tokens with no token string, only PoS, lemma etc. For example, *du* should be split into *de* and *le*, but in some instances the token string for *le* is missing. These missing tokens were omitted from this experiment.

The French Dundee Corpus does not come with a training-development-test split. We use a similar approach as for English, with the first 80% of the tokens for training, the next 10% of the tokens for development and the last 10% for testing. No sentences were split into separate sets. That results in 43,383 tokens in 1,585 sentences for training, 5,407 tokens in 240 sentences for development, and 5,444 tokens in 178 sentences for testing.

The tagset of the French Treebank was automatically mapped to the Universal PoS tag set (Petrov et al., 2011). We make the aligned, morphosyntactic annotation for the French Dundee Corpus available at `https://bitbucket.org/lowlands/release`.

2.3 Reading differences between English and French

This section discusses the results of existing studies comparing reading in French and English. The two main studies used the two Dundee corpora for their analysis.

Pynte and Kennedy (2006) compared the eye movements of the French and English Dundee corpus to explore local effects (e.g., word frequency, word length, local context) and global effects (e.g., predictability, reading strategy, inspection strategy) on five eye movement metrics.

They first of all noted that French was read slower than English with more and longer fixations. This effect is significant and is even more pronounced for long words and there are also significantly more re-fixations for French compared to English. Kennedy and Pynte (2005) argue that re-fixations reflect the most crucial difference between French and English. Besides being an obvious difference in the processing of the target word, more re-fixations also enhance preview of the next word. Pynte and Kennedy (2006) report that participants of the English and French experiments were matched (though not on which factors) and that the procedure, including calibration technique, equipment, control software, instructions, and data-reduction software, were identical across language, though the French data was collected in Aix-en-Provence, France and the English data in Dundee, UK. Therefore they ascribed this difference to the text itself. Even though they found that French words (5.2 characters) are on average longer than English ones (4.7 characters), there are more two-letter words in French (19.7%) than in English (17.2%). Therefore Kennedy and Pynte (2005) suggest that the reading difference is due the distribution of information across the letters of a given words, which is different across these two languages. For example, in French, terminal accents, case markers, and gender and tense marking convey crucial morphological information. This is in line with their finding that eye movements in the English part of the Dundee Corpus were more sensitive to the length of the next word, whereas French showed equivalent effects of the informativeness of the word-initial trigram.

	− suffix feats			+ suffix feats		
TR-TE	**No gaze**	**Token**	**Type**	**No gaze**	**Token**	**Type**
Development set accuracy						
EN-EN	77.44	80.01	**83.38**	80.21	81.46	**83.86**
FR-EN	**73.16**	72.92	72.92			
FR-FR	82.45	83.08	**86.55**	83.39	84.11	**87.52**
EN-FR	79.38	80.86	**80.97**			
Test set accuracy						
EN-EN	76.49	78.49*	**82.14***	80.37	80.60	**83.25***
FR-EN	71.38	71.39	**71.58**			
FR-FR	81.30	82.27*	**85.03***	83.16	83.30*	**86.22***
EN-FR	78.34	79.83*	**79.92***			

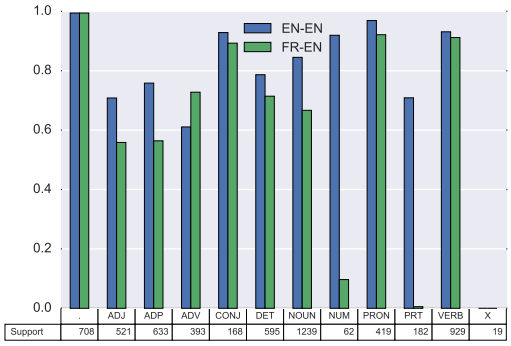
Table 1: Accuracy on development and test set for type-and token-level experiments. Best condition per experimental set-up per language combination in bold. *) For test set results: $p < 0.001$ according to mid-p McNemar test when compared to baseline.

Overall, Pynte and Kennedy (2005; 2006) conclude that the English and French inspection strategies are remarkably similar, which is the same conclusion Sparrow et al. (2003) made when testing the English EZ reader model on another eye movement corpus of 134 words of French. Kennedy and Pynte (2005) provide an analysis of the statistical differences between English and French, but besides re-fixations being more frequent in French, they seem to conclude that the reading is in many respects similar, which is also supported by their choice of mainly analyzing French and English jointly.

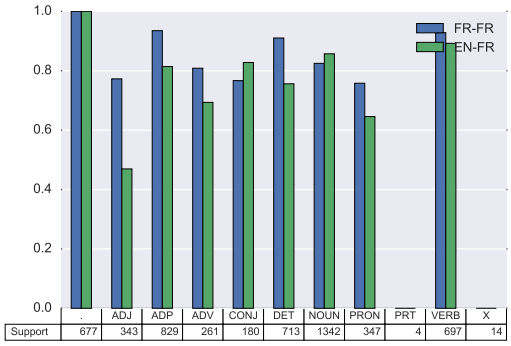
The treebank annotation includes sentence boundaries, which makes it possible to compare the length and the complexity of the sentences for both languages. We find that the average sentence length of the English training set is 24.7 tokens (SD 13.1). For French it is 28.7 tokens (SD 17.8). Sentence length was not considered by Pynte and Kennedy (2005; 2006). A consequence of longer sentences is that reading difficulty increases. The Coleman-Liau index (Coleman and Liau, 1975) is 10.38 for the English training set and 12.98 for the French.[1] This could stem from different writing styles in *Le Monde* and *The Independent* or a biased sampling of articles.

The conclusion can go no further than to say that French and English readers *can* display a more or less similar inspection strategy when reading text under matched conditions. Some effects, e.g., the fact

[1] calculated using `http://www.online-utility.org/`

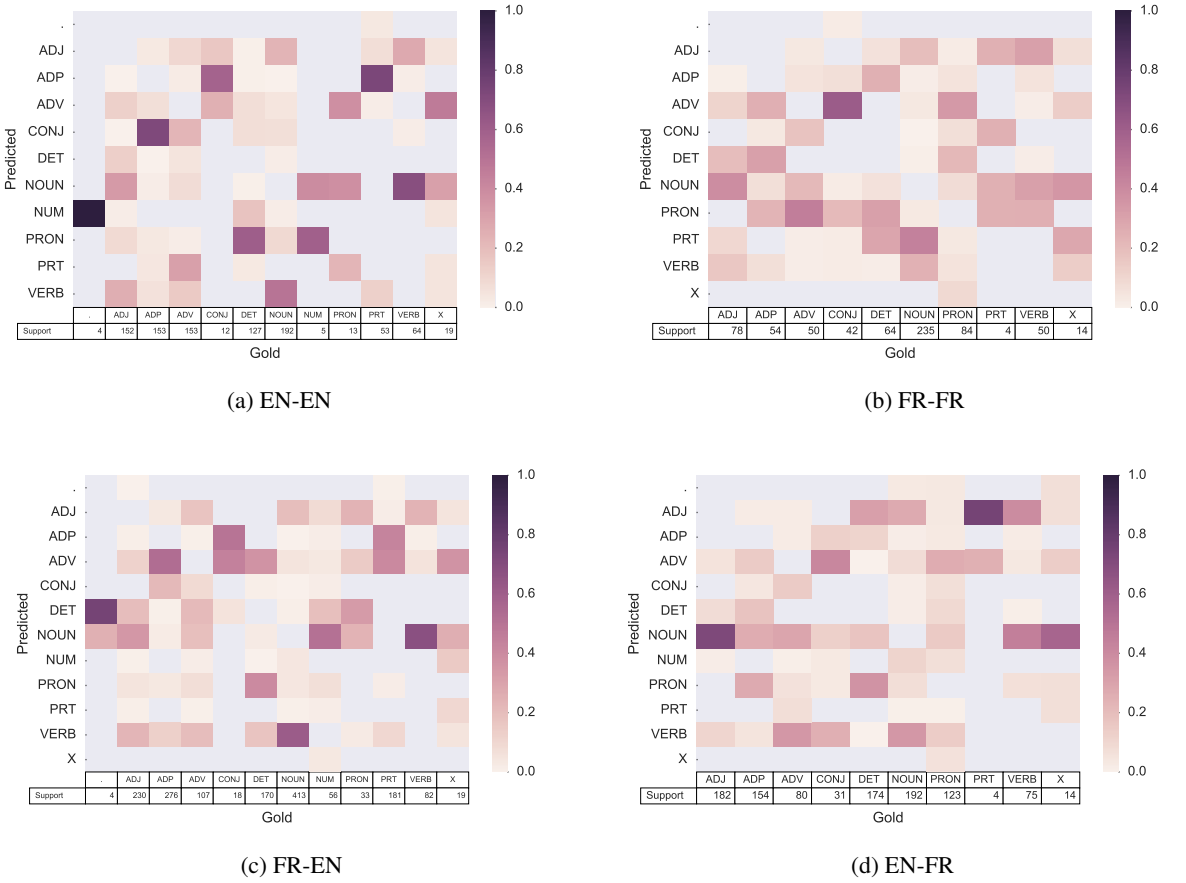


(a) EN-EN and FR-EN



(b) FR-FR and EN-FR

Figure 3: Accuracy on development set for all PoS classes.



(a) EN-EN

(b) FR-FR

(c) FR-EN

(d) EN-FR

Figure 4: Erroneous predictions per gold PoS for all combinations of training and testing language on development set.

that word-initial trigrams are more important for fixation durations in French than in English, could be due to cross-lingual differences in the spelling of the two languages, leading to re-fixations in order to increase preview. But slower reading could also be partly due to the presence of more difficult texts in the French corpus. See Section 7 for a further discussion on grammatical processing differences across languages.

2.3.1 Comparing reading of PoS for English and French

The statistics presented in the following section were computed on the French and English training sets and extends the comparison of Section 2.3 with respect to PoS. We show that the PoS classes are overall read similarly across the two languages with few exceptions due to systematic biases.

Figure 1 shows the distribution of PoS classes in the English and French data. The biggest differences are that there are no NUM tags in French. This is due to the annotation scheme and our automatic mapping, in which no tags map to NUM. There are also very few particles in the French data compared to English.

Figure 2 shows boxplots for two different reading metrics: number of fixations and first pass duration, across PoS class for English and French. The first pass duration is the sum of fixation durations for a token in the first pass through the text. This measure is said to encompass early syntactic and lexical processing. The number of fixations encompasses re-fixations and regressions to a token and reflects later syntactic and semantic processing.

Note that punctuation is almost always glued to a word and any eye movements on a punctuation will mainly—if not solely—reflect the processing of the other token. Therefore punctuation is excluded from Figure 2.

When comparing Figure 2d and 2b, we can confirm Pynte and Kennedy's (2006) finding that fixations

are generally longer in the French portion than in the English portion of Dundee. Average gaze duration in the training set is 236 ms for English and 303 ms for French.

It can be seen from Figure 2 that the measures differ across PoS for most classes in an intuitive way. For instance, PoS classes of short, frequent, closed-class words such as CONJ, ADP, PRON and DET get fewer and shorter fixations than, e.g., NOUN, VERB, ADJ, and ADV. This seems to be consistent across the two languages, and is in line with a similar analysis for English (Barrett and Søgaard, 2015a) for a smaller data set of naturally occurring text from five different domains.

The PRT category seems to be an exception. In French, PRT seems to require extensive early and late processing. Remember from Figure 1 that there are more PRTs for English (3.6%) and fewer for French (0.05%). The sets of PRT words for the two languages reveal a systematic bias in the annotation scheme or automatic mapping. For the French training set, the set of PRTs is {*vice-*, *pseudo-*, *post-*, *contre-*, *anti-*, *non-*, *quasi-*, *soviéto-*, *supra-*, *néo-*, *inter-*}. For English it is {*off*, *down*, *To*, *about*, *on*, *in*, *over*, *around*, *back*, *up*, *out*, *to*, *away*, *'*, *'s*}. French particles are therefore always at least two-token visual units that seem to be quite infrequent as well as long, whereas English particles are short and frequent.

3 Features

For our weakly supervised PoS tagging experiments, we use 22 gaze features that measure both early processing and late processing. They are equivalent to the 22 gaze features used by Barrett et al. (2016). Early processing measures are said to reflect different aspects of early syntactic and semantic processing and include first pass duration and first fixation duration. Late processing measures reflect, e.g., late syntactic and semantic integration (Rayner, 1998). Examples are number and duration of regressions going to a word, as well as the total reading time for a word.

Non-gaze features are usually included in eye movement models, because they explain a lot of the variance in fixation durations. Word frequency and word length together have been found to explain 69% of the variance in the mean gaze duration (Carpenter and Just, 1983). Like Barrett et al. (2016), we use word length, log word frequencies from a big corpus and log word frequencies from the Dundee training set for the target word, and the previous and next words. From the Dundee training set, we also extract the forward and backward transitional probability, i.e., the conditional probabilities for a word given the next or previous word. Our non-gaze features are almost equivalent to Barrett et al. (2016). The only difference is that they also used forward and backward transitional probabilities from a big corpus.

The big corpus log frequencies were obtained from the British National Corpus[2] for English, extracted with KenLM (Heafield, 2011) and Lexique[3] for French. The Dundee log frequencies were calculated on the respective training sets using CMU Language Modeling Toolkit[4] with Witten-Bell smooting.

In total we have 29 features. All features are first averaged over all 10 readers of the corpus, then scaled to a value between 0 and 1 by minmax scaling. The best model of the feature ablation study of Barrett et al. (2016) used all features, which suggests that grammatical processing of a broad set of PoS categories is reflected across many features and need non-gaze features as well.

4 Experiment

We replicate the experimental setup of Barrett et al. (2016), which used the best model from Li et al. (2012), a second-order hidden Markov model with maximum entropy emissions (SHMM-ME) constrained by Wiktionary tags such that emissions are confined to the allowed PoS tags of the Wiktionary given that the token exists in the Wiktionary. Li et al. (2012) report considerable improvements from the Wiktionary contraint when comparing to unsupervised methods.

The second-order model includes transition probabilities from the antecedent state like a first order model (Berg-Kirkpatrick et al., 2010) as well as from the second-order antecedent state.

We use the original implementation of Li et al. and we also include a subset of their word-level features, viz., four features detecting hyphens, numerals, punctuation and capitalization. We leave out the three

[2] http://www.natcorp.ox.ac.uk

[3] http://www.lexique.org

[4] http://www.speech.cs.cmu.edu/SLM/toolkit.html

suffix features from Li et al.'s basic feature model, as these features do not transfer across languages. These features were also included by Barrett et al. (2016).

We use the English Wiktionary dumps made available by Li et al.[5] The French Wiktionary dump is from Wisniewski et al. (2014) and does not include any punctuation. We therefore augment it with all punctuation entries from the English Wiktionary. Furthermore, tokens for the tag ADP are completely missing from the French Wiktionary, and the tokens for the class DET were sparse. We therefore add all examples of DET and ADP from the French training set to the French Wiktionary.

For the cross-lingual experiments, we use the union of the French and the English Wiktionary dictionaries.

Barrett et al. (2016) used Li et al.'s model for weakly supervising PoS induction with gaze features for English, and performed model tuning and feature ablation. We use their best hyper-parameter setting, i.e., five EM iterations, as well as the best feature combination: all features. Following Barrett et al. (2016), we try token-level and type-level features. For the token-level experiments, each token is represented by its feature vector. For the type-level experiments, each token is represented by an average of the feature vectors for all occurrences of the lower-cased word type of the training set.

5 Results

The tagging accuracy for all combinations of training and testing language on the development set and the test set can be seen in Table 1.

For all conditions, type-level features work better than token-level, though the type-level improvement over the baseline is not significant for FR-EN.

The English monolingual condition plus suffix is almost equivalent to the best model in Barrett et al. (2016). The only difference is the two missing non-gaze features described in Section 3. On the test set, they report a baseline accuracy of 79.77, a token-level accuracy of 81.00, and a type-level accuracy of 82.44, which is in line with our results. We observe that the suffix features seem to help in the monolingual conditions. For monolingual conditions, we confirm that type-level gaze-features and token-level ones outperform the baseline. These differences are significant, except for the EN-EN token-level plus suffix condition.

FR-FR PoS tagging seems to be a slightly an easier task than EN-EN PoS tagging, achieving overall higher accuracies.

The cross-lingual conditions generally achieve lower performance than the monolingual. When training on English and testing on French, both token-level and type-level conditions are significantly better than baseline.

Metric	Cosine sim
n refixations	0.6318
First pass duration	0.8480
Re-read probability	0.8489
n fixations	0.9097
Total fixation duration	0.9217
n regressions to	0.9354
n long regressions from	0.9375
Total duration of regressions from	0.9377
Total duration of regression to	0.9385
n regressions from	0.9404
n long regressions to	0.9644
Fixation probability	0.9795
w-1 fixation duration	0.9839
w+1 fixation duration	0.9934
w-1 fixation probability	0.9947
w+2 fixation duration	0.9961
w+1 fixation probability	0.9967
w-2 fixation probability	0.9975
w+2 fixation probability	0.9986
First fixation duration	0.9992
Mean fixation duration	0.9992
w-2 fixation duration	0.9992

Table 2: Cosine similarity between PoS averaged French and English train set gaze vectors across gaze features. Sorted by similarity.

6 Error Analysis

There are—as expected—more errors when using cross-lingual gaze data. This section will explore these errors by comparing the predictions of the cross-lingual experiments with the predictions of the mono-

[5] https://code.google.com/archive/p/wikily-supervised-pos-tagger/

lingual experiments. All analysis is on the development set output of the type-level models. We compare them to the output of the type-level monolingual models.

Figure 3 shows accuracy scores per PoS class comparing experiments with same test set. The accuracy of punctuations is due to the basic feature model and the Wiktionary constraints—not the eye movement measures. PRT and NUM are real challenges for FR-EN compared to EN-EN. This can be assumed to be due to the different use of the PRT tag and the missing NUM class in the French dataset described in Section 2.3.1. ADJ also seems like a cross-lingual challenge, though harder when trained on English and tested on French than the other way around.

Figure 4 shows the erroneous predictions per gold PoS tag, allowing us to compare error types across experiments. When comparing Figure 4a and Figure 4c, both evaluated on English, most classes seem to have almost the same set of misclassified labels though for some labels in different magnitude or ratio depending on whether they are trained on English or French. The main differences are: when trained on French, ADP and ADJ are generally more often misclassified, ADP is not mainly misclassified as CONJ, but more often as ADV, DET is also misclassified as VERB and ADV, PRT is misclassified as ADV and not mainly as ADP.

When comparing Figure 4b and Figure 4d, both evaluating on French, we also find that for many of the PoS classes, the misclassifications are of the same type, though different in magnitude or ratio. The main differences we observe when training on English are: ADJ is mainly misclassified as NOUN instead of ADP, ADV, DET, NOUN, and PRT; ADV is misclassified as VERB; DET is never misclassified as PRT, but more often as NOUN and ADJ; and NOUN is rarely misclassified as PRT. The last error probably has to do with the long gaze durations for PRT in the French data (resembling gaze durations of NOUNs) opposed to the short gaze durations of English PRT.

Table 2 shows the cosine similarity between the English and French PoS-averaged gaze vectors from the train set for all gaze features. This gives information about which gaze feature averages differ between French and English PoS. Pynte and Kennedy (2006) found that French had more re-fixations than English, which is reflected in the table. Measures correlating with re-fixations like re-read probability, number of fixtions, and total fixation duration are naturally also different across languages. First pass duration is not directly correlated with number of re-fixations, and must be considered an distinct pattern.

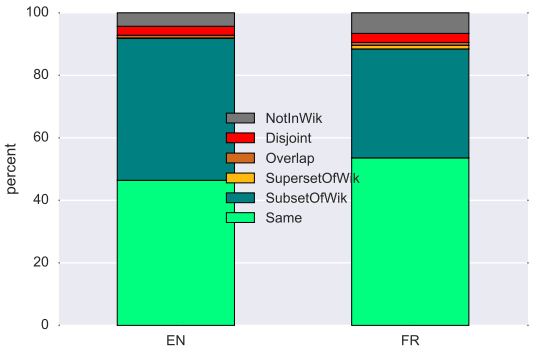


Figure 5: Development set word type lookup in Wiktionary for English and French: the percentage of word types assigned a set of tags that is either: identical to, a subset of, a superset of, overlapping with, disjoint with, or not in the Wiktionary.

6.1 Wiktionary agreement

Figure 5 shows the word types for the English and French development set according to their representation in the respective monolingual Wiktionary. This figure is inspired by Li et al. (2012). For English, more PoS types agree with the Wiktionary (Same or SubsetOfWik) than for French. We also computed token-level accuracies, where a tag licensed by Wiktionary counts as correct. For the French development set, this maximum dictionary accuracy is 0.95, whereas for English it is 0.92.

7 Discussion

We presented four experiments with PoS induction using gaze data in a monolingual and cross-lingual setup with a second-order hidden Markov model. Our experiments confirm the main conclusion from Barrett et al. (2016), viz., that type-level gaze vectors improve PoS induction. We replicated their result

for English and report the same finding for French as well as for French when trained on English gaze vectors.

It is difficult to determine how much the relatedness of the French and English languages is responsible for the ability of the model to generalize cross-lingually. The psycholinguistic literature does not reveal how different PoS categories are processed across languages; most experimental work in the literature studies single phenomena in one language. For instance, in reaction time studies of lexical decision tasks it has been found that the processing of English plural and singular nouns is influenced by surface frequency only[6] (Sereno and Jongman, 1997), whereas for Dutch (Baayen et al., 1997) and French (New et al., 2004), the lexical processing of singular and plural nouns is influenced by the base frequency[7]. The English data thus support a full-storage cognitive model, whereas the French and the Dutch data support the Parallel Dual-Route model where a word is processed as segments in parallel with whole word processing. These results suggest that nouns are processed differently in the brain for native speakers of different languages. This means that our results may not generalize to other combinations of languages and in the specific case of nouns it suggests that Dutch and French nouns are processed more similarly than French and English.

8 Conclusion

This is, to the best of our knowledge, the first study to explore whether gaze features generalize from one language to another for a broad set of syntactic categories. We used a type-constrained second-order HMM for monolingual and cross-lingual PoS induction on the English and French portions of the Dundee eye tracking corpus. We experimented with both token-level and type-level features and confirmed that type-level gaze features improve monolingual PoS induction for both English and French. We also showed that type-level gaze features significantly improve PoS induction for French, even when the model is trained on English gaze vectors.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In *Treebanks*, pages 165–187.

Harald R Baayen, Ton Dijkstra, and Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37(1):94–117.

Maria Barrett and Anders Søgaard. 2015a. Reading behavior predicts syntactic categories. *CoNLL 2015*, pages 345–349.

Maria Barrett and Anders Søgaard. 2015b. Using reading behavior to predict grammatical functions. In *Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pages 1–5.

Maria Barrett, Željko Agić, and Anders Søgaard. 2015. The Dundee treebank. In *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 242–248.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *ACL*, pages 579–584.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Cote, John DeNero, , and Dan Klein. 2010. Painless unsupervised learning with features. In *NAACL*, pages 582–590.

Patricia A Carpenter and Marcel Adam Just. 1983. What your eyes do while your mind is reading. *Eye movements in reading: Perceptual and language processes*, pages 275–307.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197.

[6] the token frequency of a word form

[7] the sum of the frequencies of all inflections of a word

Alan Kennedy and Joël Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision research*, 45(2):153–168.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. *Poster presented at the 12th European Conference on Eye Movement*.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *NAACL*, pages 1528–1533.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *EMNLP*, pages 1389–1398.

Boris New, Marc Brysbaert, Juan Segui, Ludovic Ferrand, and Kathleen Rastle. 2004. The processing of singular and plural nouns in French and English. *Journal of Memory and Language*, 51(4):568–585.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. CoRR abs/1104.2086.

Joel Pynte and Alan Kennedy. 2006. An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. *Vision Research*, 46(22):3786–3801.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422.

Joan A Sereno and Allard Jongman. 1997. Processing of English inflectional morphology. *Memory & Cognition*, 25(4):425–437.

Laurent Sparrow, Sébastien Miellet, and Yann Coello. 2003. The effects of frequency and predictability on eye fixations in reading: An evaluation of the EZ reader model. *Behavioral and Brain Sciences*, 26(04):503–505.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *EMNLP*, volume 14, pages 1779–1785.